Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Adding attention to the neural ordinary differential equation for spatio-temporal prediction

Peixiao Wang, Tong Zhang, Hengcai Zhang, Shifen Cheng & Wangshu Wang

To cite this article: Peixiao Wang, Tong Zhang, Hengcai Zhang, Shifen Cheng & Wangshu Wang (07 Nov 2023): Adding attention to the neural ordinary differential equation for spatio-temporal prediction, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2023.2275160

To link to this article: https://doi.org/10.1080/13658816.2023.2275160

View supplementary material

Published online: 07 Nov 2023.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgis20

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2023.2275160

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Adding attention to the neural ordinary differential equation for spatio-temporal prediction

Peixiao Wang[a,b,c], Tong Zhang[b], Hengcai Zhang[a,c], Shifen Cheng[a] and Wangshu Wang[d]

[a]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [b]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [c]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China; [d]Research Unit Cartography, Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria

ABSTRACT

Explainable spatio-temporal prediction gains attraction in the development of geospatial artificial intelligence. The neural ordinal differential equation (NODE) emerges as a new solution for explainable spatio-temporal prediction. However, challenges still need to be solved in most existing NODE-based prediction models, such as difficulty modeling spatial data and mining long-term temporal dependencies in data. In this study, we propose a spatio-temporal attentional NODE (STA-ODE) to address the two challenges above. First, we define a spatio-temporal ordinary differential equation to predict a value at each time iteratively by a novel spatio-temporal derivative network. Second, we develop an attention mechanism to fuse multiple prediction values for capturing long-term temporal dependencies in data. To train the STA-ODE model, we design a loss function that aligns the prediction results in spatial dimension with prediction results in temporal dimension to calibrate the parameters of the model. The proposed model was validated with three real-world spatio-temporal datasets (traffic flow dataset, PM2.5 monitoring dataset, and temperature monitoring dataset). Experimental results showed that STA-ODE outperformed seven existing baselines regarding prediction accuracy. In addition, we used visualization to demonstrate the sound interpretability and prediction accuracy of the STA-ODE model.

ARTICLE HISTORY
Received 22 April 2023
Accepted 20 October 2023

KEYWORDS
Geospatial artificial intelligence; spatio-temporal prediction; spatio-temporal attention; neural ordinary differential equation

1. Introduction

Spatio-temporal prediction, which relies on spatio-temporal data to predict the unknown system states in time and space (Janowicz *et al*. 2020, Xie *et al*. 2020, Xu *et al*. 2021), is a hot research topic in geospatial artificial intelligence (GeoAI). In recent

CONTACT Tong Zhang zhangt@whu.edu.cn

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13658816.2023.2275160.

© 2023 Informa UK Limited, trading as Taylor & Francis Group

years, with the rapid development of the Internet of Things, a large amount of sensor data has been collected to provide critical data support for spatio-temporal prediction (Kang *et al*. 2022, Wang, Zhang, and Hu 2022). At present, spatio-temporal prediction technology has been widely used in intelligent transportation, weather forecasting, earthquake early warning, and other applications (Cheng *et al*. 2021, Wang et al. 2022b).

Existing spatio-temporal prediction models can generally be categorized as knowledge-driven or data-driven spatio-temporal models, each of which has its own set of strengths and weaknesses (Wang *et al.* 2022a, Zheng *et al*. 2023). Specifically, knowledge-driven models establish specific mathematical equations to describe complex spatio-temporal patterns based on prior knowledge, resulting in high interpretability but poor predictive performance (Cheng *et al*. 2020, Li *et al*. 2021). Unlike knowledge-driven models, data-driven models tend to ignore the prior knowledge accumulated by previous scholars. Instead, they establish machine learning or deep learning models to automatically mine the complex spatio-temporal patterns from data. Although data-driven models may achieve superior prediction accuracy, they are often regarded as black-box models with poor interpretability (Liu *et al*. 2016, Li *et al*. 2020, Zhang *et al*. 2020). Most existing spatio-temporal prediction models are still struggling to balance prediction accuracy with interpretability (Janowicz *et al*. 2020, Sagi and Rokach 2020).

In recent years, the increasingly popular neural ordinal differential equation (NODE) model has provided a novel appoach to interpretable spatio-temporal prediction (Chen *et al*. 2018). By establishing a connection between deep learning and ordinary differential equations (ODEs) using a neural network parameterized derivative model, NODE can predict the unknown state of the spatio-temporal system (Ji *et al*. 2022). Specifically, in NODE, the prediction value is defined as the solution to an initial value problem for an ODE at a given time, which is then iteratively solved for each time step using the derivative network. Since the solution of NODE has an explicit mathematical expression and physical interpretation (discussed further in Section 3.2), the NODE-based model can be regarded as an explainable prediction model. At present, many scholars have proposed various NODE-based prediction models that have achieved high prediction accuracy and interpretability, such as spatio-temporal ordinary differential equations (ST-ODEs) (Zhou *et al*. 2021), spatio-temporal graph ordinary differential equations (STG-ODEs) (Fang *et al*. 2021), recurrent neural network ordinary differential equations (ODE-RNNs) (Rubanova *et al*. 2019), and long short-term memory ordinary differential equations (ODE-LSTMs) (Lechner and Hasani 2020). Although the above-mentioned NODE-based prediction models try to balance the prediction accuracy and interpretability, they are still inadequate. First, the derivative network, the core component of the NODE model, only models temporal information rather than spatial information, hindering NODE-based models'ability to mine spatial dependencies in data. Second, the prediction value of NODE-based models heavily depends on the initial value of the ODE, making it difficult for NODE-based models to capture long-term temporal dependencies in data. In other words, prediction results are influenced by recent observations and those from the distant past.

In general, although the above-mentioned NODE-based prediction models improve the interpretability of data-driven models, it is still challenging to obtain state-of-the-art accuracy for spatio-temporal prediction tasks. Therefore, we introduce a novel spatio-temporal attentional neural ordinary differential equation (STA-ODE), which offers state-of-the-art predictive accuracy and reasonable interpretation. The main contributions of this study are summarized as follows:

1. We define a spatio-temporal derivative network to assist the STA-ODE in mining spatio-temporal correlations in data by accounting for both temporal and spatial location information. With the spatio-temporal derivative network, the STA-ODE model can solve the prediction value iteratively at multiple times in an interpretable manner.
2. We propose a novel attention mechanism to fuse multiple prediction values and-capture long-term temporal dependencies in data. In addition, inspired by multi-view learning, we designed a loss function that aligns the prediction results in spatial dimension with prediction results in temporal dimension to calibrate the parameters of the model.
3. We evaluated the prediction performance of the STA-ODE model using three actual spatio-temporal datasets (ie traffic volume dataset, PM2.5 monitoring dataset, and temperature monitoring dataset). The results proved the advantages of our model over seven baseline methods. In addition, we used visualization to demonstrate the sound interpretability and prediction accuracy of the STA-ODE model.

2. Literature review

In this subsection, we first review knowledge-driven spatio-temporal prediction models, then review data-driven spatio-temporal prediction models, and lastly review the NODE-based spatio-temporal prediction models that have emerged in recent years.

2.1. Knowledge-driven spatio-temporal prediction models

Knowledge-driven spatio-temporal prediction models are a type of prediction method that assumes that spatio-temporal data adheres to definite physical laws and mechanisms in either the spatial or temporal dimension. These models establish specific mathematical equations to describe spatio-temporal data patterns (Huang *et al*. 2021). For example, the inverse distance weighting (IDW) model assumes that the spatial distribution of data conforms to the first law of geography and predicts future spatio-temporal data by computing the distance between the prediction location and the observed location (Bartier and Keller 1996). Kriging interpolation assumes that the spatial distribution of the spatio-temporal data satisfies the second-order stability and uses a covariance function to obtain an optimal, linear, and unbiased estimation of the prediction data (Pesquer *et al*. 2011). Autoregressive integrated moving average (ARIMA) assumes that the spatio-temporal data satisfies the time stationarity in the time dimension and infers the future spatio-temporal data based on the historical data from several preceding moments (Yozgatligil *et al*. 2013). At the same time, several researchers have proposed advanced prediction statistical models based on the above models, including spatio-temporal inverse distance weighting (ST-IDW) (Li *et al.* 2014), spatio-temporal kriging (ST-Kriging) (Aryaputera *et al.* 2015), and spatio-temporal autoregressive integrated moving average (ST-ARIMA) models (Peibo Duan *et al.* 2016). Knowledge-driven models establish specific mathematical equations that describe geographical phenomena, resulting in excellent model interpretability but poor prediction accuracy in spatio-temporal prediction tasks. The reason is that the knowledge-driven methods rely on strict mathematical assumptions, which are challenging to meet in the actual geographical environment.

2.2. Data-driven spatio-temporal prediction models

In recent years, with the rapid development of artificial intelligence and high-performance computing, data-driven models have gradually become the mainstream models for spatio-temporal prediction (Ermagun and Levinson 2018). Compared to knowledge-driven spatio-temporal prediction methods, data-driven spatio-temporal prediction methods do not require datasets to obey specific mathematical laws, but instead train a supervised machine learning to establish a functional mapping between input data and output data for spatio-temporal prediction tasks (Xu *et al.* 2021, Wang *et al.* 2023), such as k-nearest neighbors (Yu *et al.* 2016, Cheng *et al.* 2018), tensor factorization model (Yu *et al.* 2016, Chen and Sun 2022), spatio-temporal residual networks (Zhang *et al.* 2017), and generative adversarial neural networks (Zhu *et al.* 2020a). In addition, considering non-Euclidean structures of datasets, the graph convolutional neural network (GCN) is also used for spatio-temporal prediction tasks and further improves the prediction accuracy, such as temporal graph convolutional network (T-GCN) (Zhao *et al.* 2020), spatio-temporal graph convolutional network (ST-GCN) (Yu *et al.* 2018), graph attention temporal convolutional network (GATCN) (Zhang *et al.* 2021), and Place GCN (Zhu *et al.* 2020b). Compared to the knowledge-driven spatio-temporal prediction methods, although the data-driven models have greatly improved the prediction accuracy, their inherent black-box characteristics lead to poor interpretability of the models (Samek *et al.* 2021). In the field of geosciences, the interpretability and transparency of models are one of the top priorities of GeoAI (Janowicz *et al.* 2020). However, most existing data-driven models cannot balance the prediction accuracy and interpretability.

2.3. NODE-based spatio-temporal prediction models

In order to achieve both high prediction accuracy and explainability of prediction models, hybrid approaches that integrate data-driven and knowledge-driven models have been developed. One such model is NODE, recognized as one of such representative models (Chen *et al.* 2018). The NODE model uses a derivative network to establish the relationship between deep learning and ordinary differential equations, resulting in good prediction accuracy and interpretability. At present, several scholars have proposed a series of variant NODE-based models to solve spatio-temporal prediction tasks. For example, Rubanova *et al.* (2019) proposed the ODE-RNNs model by combining the recurrent neural network with NODE and tested it on a toy dataset of

1000 periodic trajectories. Zhou *et al.* (2021) proposed the spatio-temporal ordinary differential equations (ST-ODEs) and applied them to traffic flow prediction tasks. Huang *et al.* (2021) and Fang *et al.* (2021) extended the classic NODE to accommodate the graph structure, and thus proposed graph-based ordinary differential equations, ie graph ODEs. Ji *et al.* (2022) integrated the physical mechanism of traffic flow mechanics into NODE and developed a spatio-temporal differential equation network (STDEN). Compared to purely data-driven models, although these NODE-based prediction models significantly improve the interpretability through differential equations, there are still two shortcomings. On the one hand, these NODE-based prediction models face challenges in mining the spatial dependencies in the spatio-temporal data. On the other hand, it is challenging for the above NODE-based prediction models to discover long-term temporal dependencies in data.

Therefore, we propose a novel spatio-temporal prediction model called STA-ODE with state-of-the-art prediction accuracy and excellent interpretation. The STA-ODE mode extends the traditional temporal derivative network to a spatio-temporal derivative network, which helps STA-ODE to discover spatio-temporal correlations in data. Additionally, we integrate an attention mechanism into the STA-ODE model to capture long-term temporal dependencies in data.

3. Preliminaries

3.1. Problem definitions

Graph structure is a typical data structure, and many spatio-temporal prediction tasks can be abstracted as graph-based problems. Therefore, we build the STA-ODE model based on the graph structure. Before introducing the details of the STA-ODE model, we present the relevant definitions required by the STA-ODE model and the mathematical description of the prediction problem. In the appendix, we also list necessary notations and corresponding illustrations (see Table S1 and Figure S1 for details).

Definition 1 (Graph), As shown in Figure 1, the graph $G =< V, E >$ represents the graph structure extracted from the monitoring sites, where $V = \{v_i\}_{i=1}^{N}$ represents N nodes in G, ie N monitoring sites; $E = \{e_{ij}\}$ represents the relationships between node v_i and node v_j. It is noted that the monitoring sites in the study area are abstracted as a fully connected graph, and the connection weights (ie relationships) between nodes are learned by the model.

Definition 2 (Spatio-temporal State), The spatio-temporal state $x_t^i \in \mathcal{R}^{1\times1}$ represents the monitored value of node v_i during time window t, such as the traffic flow or air quality per unit time (we mainly focus on a single spatio-temporal parameter monitored by monitoring sites). The spatio-temporal state of all nodes in all time windows can be expressed as a spatio-temporal state matrix $\boldsymbol{X} \in \mathcal{R}^{N\times T}$, where $\boldsymbol{x}^i = \left\{x_t^i\right\}_{t=1}^{T} \in \mathcal{R}^{T\times1}$ represents the time series of node v_i in all time windows, $\boldsymbol{x}_t = \left\{x_t^i\right\}_{i=1}^{N} \in \mathcal{R}^{N\times1}$ represents the spatial sequence of all nodes in the time window t, N represents the number of nodes, and T represents the total number of all time windows.

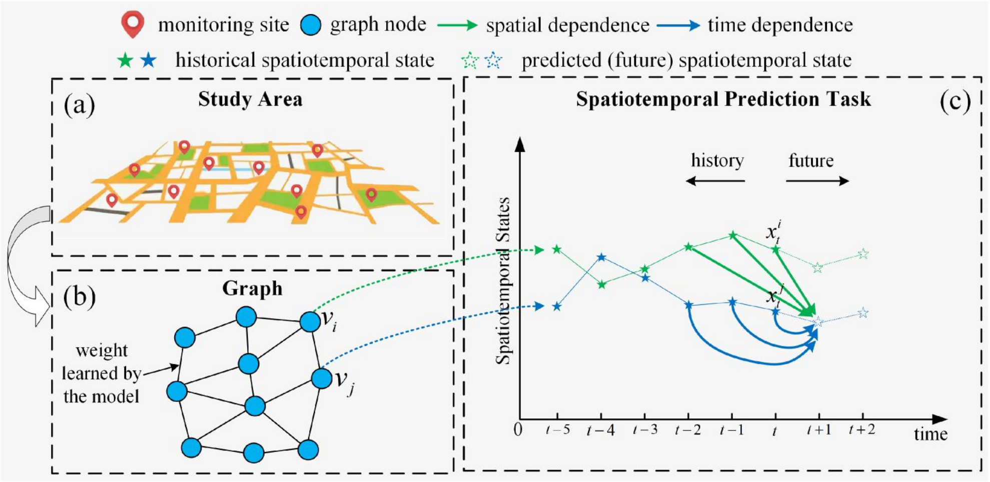

Figure 1. Illustration of definitions: (a) monitoring sites in the study area, (b) the graph structure derived from the monitoring sites, and (c) the spatio-temporal prediction task.

As shown in Figure 1, this study aims to build a functional model $\mathcal{F}(\cdot)$ for predicting future spatio-temporal data based on the graph structure G and the spatio-temporal state matrix $\boldsymbol{X}$. Specifically, the process is shown in Equation (1),

$$\hat{\boldsymbol{x}}_{t+1} = \mathcal{F}\left(\boldsymbol{X}_{t-k+1}^{t}, G; \boldsymbol{W}\right) = \mathcal{F}\left(\{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t}\}, G; \boldsymbol{W}\right) \tag{1}$$

where $\boldsymbol{X}_{t-k+1}^{t} = \{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t}\} \in \mathcal{R}^{n\times k}$ represents the historical data required by the prediction model, and k stands for the time dependent step; G represents the graph structure abstracted by the study area; $\hat{\boldsymbol{x}}_{t+1}$ represents predicted spatio-temporal data. In this study, we focus on single-step prediction (one-step prediction) rather than multi-step prediction (prediction step greater than 1); $\mathcal{F}(\cdot)$ represents the prediction model, ie STA-ODE model; $\boldsymbol{W}$ indicates the learnable parameters in the model.

3.2. Neural ordinary differential equations

The NODE model is a time-series prediction model in the continuous time domain, disigned for time-series modeling of a single monitoring site (Chen *et al.* 2018). Specifically, NODE regards the observation value at each time of a single monitoring site as the solution of the ordinary differential equation and iteratively computes the prediction values based on the derivative network. Assuming that the time series $\boldsymbol{x}^i = \{x_t^i\}_{t=1}^{T} \in \mathcal{R}^{T\times 1}$ represents the discrete sampling of node v_i in the continuous time domain $x^i(t)$, the prediction value of node v_i at a specific time is shown in Equation (2).

$$x^i(t) = x^i(0) + \int_0^t \frac{dx^i(\tau)}{d\tau} d\tau = x^i(0) + \int_0^t g(x^i(\tau), \tau) d\tau \tag{2}$$

where $x^i(t) \in \mathcal{R}^{1\times 1}$ represents the predicted value of the monitoring site v_i at time t; $x^i(0)$ represents the initial value of NODE, which is the observed value of monitoring site v_i at time 0; and $g\left(x^i(t), t\right) = \frac{dx^i(t)}{dt}$ represents the derivative of the function $x^i(t)$ with respect to time t in the continuous time domain, which is parameterized by a

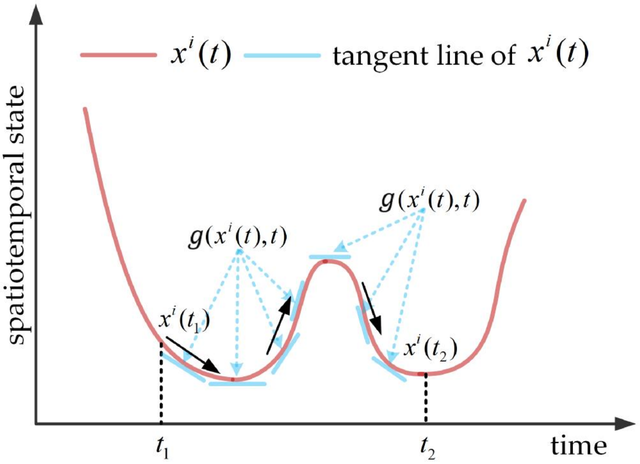

Figure 2. Illustration of the NODE-based prediction models.

neural network, namely the derivative network. As shown in Figure 2, the prediction value $x^i(t_2)$ can be obtained through multiple uphill and downhill processes based on the derivative network $g(x^i(t), t)$ based on the initial value $x^i(t_1)$. Since the NODE-based prediction models have a clear mathematical expression, they have strong interpretability.

4. Methodology

As shown in Figure 3, the STA-ODE model mainly consists of two modules: the spatio-temporal ordinary differential equation (ST-ODE) module and the spatio-temporal attention (STA) module. The ST-ODE module adapts the traditional NODE-based time-series prediction model into a spatio-temporal prediction model, while the STA module assists the ST-ODE module in capturing long-term temporal dependencies in the data. Specifically, in the ST-ODE module, we extend the traditional temporal derivative network to the spatio-temporal derivative network for solving the hidden state (similar to the hidden states in RNN) at each time (as discussed in Sections 4.1.1 and 4.1.2, respectively). In the STA module, we use spatial attention (SA) and temporal attention (TA) to fuse the hidden states of multiple times to capture the long-term temporal dependencies in data (as discussed in Section 4.1.3).

4.1. Construction of the STA-ODE

The proposed STA-ODE model addresses two challenges in most existing NODE-based prediction models, namely the challenges associated with modeling spatial data and mining long-term temporal dependencies in data. In this subsection, we introduce the implementation details of the STA-ODE model. Firstly, we define a hidden state for each graph node to improve the nonlinear fitting ability of the STA-ODE model. Secondly, we establish a spatio-temporal derivative network for the hidden states to assist the ST-ODE in modeling spatio-temporal data (ie defining the derivative of the hidden states). Thirdly, based on the defined spatio-temporal derivative network, we use the differential

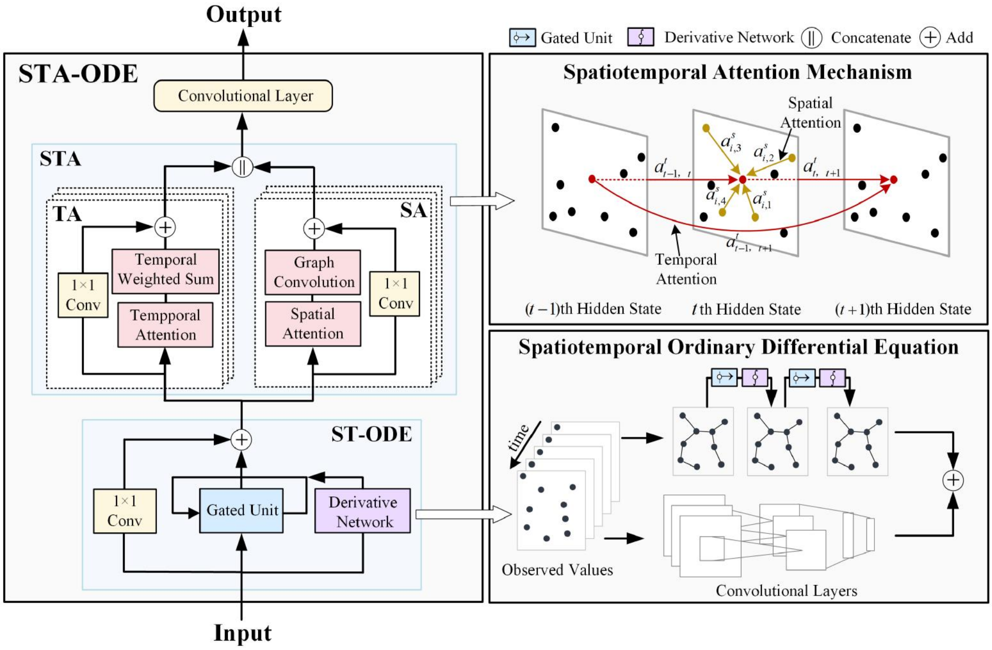

Figure 3. Overall schematic of the STA-ODE model: spatio-temporal data is used as input for the ST-ODE to obtain hidden states at multiple times for all nodes. Then, multiple hidden states are used as the input of the STA module to obtain the final prediction results.

equation to solve the hidden state of each node iteratively over time (ie forward propagation of the ST-ODE). After solving the hidden states, we integrate the spatio-temporal attention mechanism into the STA-ODE model to fuse multiple hidden states, thereby capturing long-term temporal dependencies in data (ie forward propagation of the STA).

Based on the above ideas, the forward propagation of the STA-ODE model can be roughly divided into three steps: definition of the hidden state derivative, forward propagation of the ST-ODE, and forward propagation of the STA (as discussed in Sections 4.1.1–4.1.3, respectively). Taking the spatio-temporal state matrix $\boldsymbol{X}_{t-k+1}^{t} = \{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_t\}$ as an example, the forward propagation process of the STA-ODE model can be defined by the Equation (3).

$$\begin{cases} \mathcal{H} = STODE\left(\{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_t\}, G, DN_{ST}; \boldsymbol{W}_{STODE}\right) \\ \{\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}}, \hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}}\} = STA(\mathcal{H}, G; \boldsymbol{W}_{STA}) \\ \hat{\boldsymbol{x}}_{t+1} = \left[\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} \middle|\middle| \hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}}\right] \boldsymbol{W}_o \end{cases} \tag{3}$$

where $\boldsymbol{x}_{t-k+1}$, $\boldsymbol{x}_{t-k+2}$, and $\boldsymbol{x}_t \in \mathcal{R}^{N\times 1}$ represent the observation data of graph nodes from k historical times; DN_{ST} represents the spatio-temporal derivative network of hidden states; G represents the graph structure abstracted by the study area; $STODE$ refers to the ST-ODE module, which is used to solve the hidden state of graph nodes at k historical times; $\mathcal{H} \in \mathcal{R}^{N\times d_h \times k}$ represents k hidden states obtained from the ST-ODE module, where d_h represents the dimension of the hidden state; STA represents the STA module, which is used to fuse the hidden states of k historical time steps; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}} \in \mathcal{R}^{N\times 1}$ represents the fusion values in the temporal dimension; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} \in \mathcal{R}^{N\times 1}$

represents the fusion values in the spatial dimension; $\hat{\boldsymbol{x}}_{t+1} \in \mathcal{R}^{N\times 1}$ represents the predicted values of the STA-ODE model; $\boldsymbol{W}_{STODE}$ represents the learnable parameters in the ST-ODE module; $\boldsymbol{W}_{STA}$ represents the learnable parameters in the STA module; and $\boldsymbol{W}^{o} \in \mathcal{R}^{2\times 1}$ represents the learnable parameters in the model output process.

4.1.1. Definition of hidden state derivatives

The derivative network is the core component of the NODE-based model, and its definition is crucial for extending time-series models to spatio-temporal prediction models (ie extending NODE to ST-ODE). Since the derivative networks of existing NODE-based models solely model temporal information without any spatial information, existing NODE-based models often perform poorly in spatio-temporal prediction tasks. Figure 4 further shows the reasons for the poor performance of traditional NODE-based prediction models. According to the trend of the red curve, the derivative $g\left(x^{i}(t_3), t_3\right)$ of the red curve at the time t_3 should be negative, thus suggesting that the curve $x^{i}(t)$ monitored by v_i should display a downward trend at the time t_3. Similarly, according to the trend of the grey curve, the derivative $g\left(x^{j}(t_3), t_3\right)$ of the grey curve at the time t_3 should be greater than 0, indicating an upward trend in the curve $x^{j}(t)$ monitored by v_j at time t_3. However, $x^{j}(t_3)$ is equal to $x^{i}(t_3)$, $g\left(x^{i}(t_3), t_3\right)$ will also be equal to $g\left(x^{j}(t_3), t_3\right)$ in a traditional derivative network, making it difficult to iterate the prediction values towards the ground truth. Therefore, we develop a derivative network that consider both temporal and spatial information, specifically refered to as the spatio-temporal derivative network (STDN).

Compared to the traditional derivative network, the spatio-temporal derivative network has two differences. Firstly, it solves the derivative of the hidden state rather than the derivative of the observed value, improving the nonlinear fitting ability of the STA-ODE model. Secondly, it explicitly incorporates spatial information, thereby enabling the STA-ODE model to be used for spatio-temporal prediction tasks across multiple monitoring stations instead of solely a single monitoring station. The mathematical definition of the spatio-temporal derivative network is defined in Equation (4).

$$DN_{ST} = \; g\left(\boldsymbol{H}(t), t, \{i\}_{i=1}^{N}\right) \tag{4}$$

where $\boldsymbol{H}(t) = \left\{\boldsymbol{h}^{i}(t)\right\}_{i=1}^{N} \in \mathcal{R}^{N\times d_h}$ represents the hidden states of all monitoring stations at time t, with $\boldsymbol{h}^{i}(t) \in \mathcal{R}^{1\times d_h}$ representing the hidden state of the ith monitoring

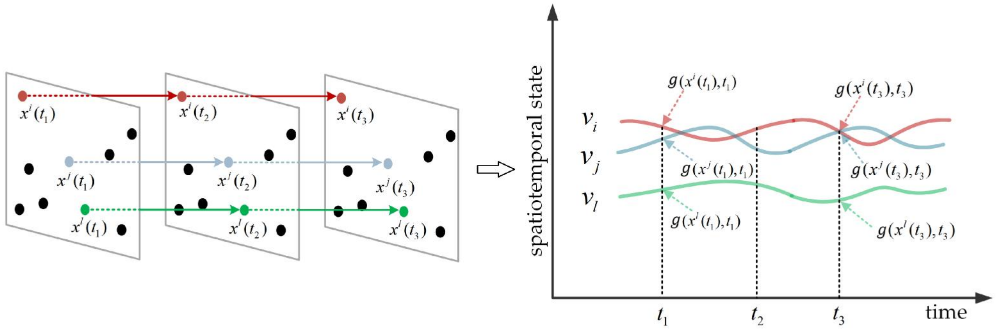

Figure 4. Parameterization of the hidden state derivative.

station at time t; $\{i\}_{i=1}^{N}$ represents the spatial location information of different monitoring stations, which is further encoded in the spatio-temporal derivative network. As shown in Figure 4, the different encoding of i and j results in distinct output values of $g(\boldsymbol{h}^{i}(t_3), t_3, i)$ and $g(\boldsymbol{h}^{j}(t_3), t_3, j)$ in the STDN, contributing to the increased prediction capability of the STA-ODE model. Similar to the derivative network in NODE (Chen *et al.* 2018), the spatio-temporal derivative network can be fitted by either a simple fully connected network or a complex full convolutional neural network (A neural network with three input variables). Generally, the complexity of neural networks plays a critical role in determining the final prediction performance of the model. In this study, we use location embedding and three-layer fully connected networks to fit the spatio-temporal derivative networks.

4.1.2. Forward propagation of the ST-ODE

After defining the spatio-temporal derivative network, the ST-ODE module is capable of modeling spatio-temporal data. The ST-ODE module can solve the hidden state in spatio-temporal data iteratively at each time using Equation (2). However, as Equation (2) is an iterative model, there is a high tendency for gradient vanishing/exploding, resulting in slow or even impossible model convergence during optimization. To alleviate the gradient vanishing or exploding during the optimization process, we propose using gating mechanisms and residual connections to accelerate the optimization efficiency of the model. Figure 5 shows the forward propagation process of the ST-ODE module, with the gating mechanism and the spatio-temporal derivative network being used to solve the hidden state at each time step. In addition, we implement a residual connection via a 1×1 convolution operation between the observation value and the hidden state due to their disparate dimensionalities. Take the spatio-temporal state matrix $\boldsymbol{X}_{t-k+1}^{t} = \{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_{t}\}$ as an example, the iterative solving process of the hidden state at each time step is shown in Equation (5), and the residual connection between

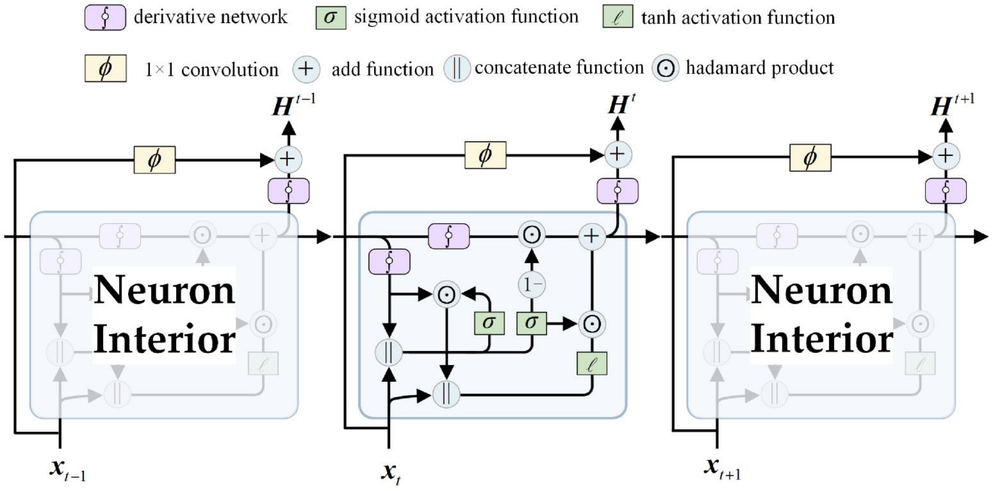

Figure 5. Forward propagation of the ST-ODE module: the hidden state of the current time is obtained through the spatio-temporal derivative network, a gating mechanism and a residual connection.

the observation value and the hidden state is shown in Equation (6).

$$\begin{cases} \boldsymbol{Z}_{t-1} = \sigma(\boldsymbol{W}_z[\boldsymbol{H}_{t-1}||\boldsymbol{x}_t]) \\ \boldsymbol{R}_{t-1} = \sigma(\boldsymbol{W}_r[\boldsymbol{H}_{t-1}||\boldsymbol{x}_t]) \\ \ddot{\boldsymbol{H}}_{t-1} = \tanh(\boldsymbol{W}_h[\boldsymbol{R}_{t-1}\odot\boldsymbol{H}_{t-1}||\ \boldsymbol{x}_t]) \\ \tilde{\boldsymbol{H}}_{t-1} = (1-\boldsymbol{Z}_{\mathrm{t}-1})\odot\boldsymbol{H}_{t-1} + (\boldsymbol{Z}_{\mathrm{t}-1}\odot\ddot{\boldsymbol{H}}_{t-1}) \\ \boldsymbol{H}_t = \tilde{\boldsymbol{H}}_{t-1} + \int_{t-1}^{t}\ g\left(\boldsymbol{H},\tau,\{i\}_{i=1}^{n}\right)d\tau \end{cases} \tag{5}$$

$$\boldsymbol{\mathcal{H}} = [\boldsymbol{H}_{t-k+1}||\ldots\ldots||\boldsymbol{H}_t] + \Phi_{\boldsymbol{W}_{ic}}*\boldsymbol{X}_{t-k+1}^{t} \tag{6}$$

where $\boldsymbol{\mathcal{H}} \in \mathcal{R}^{N\times d_h\times k}$ represents k hidden states after residual connection; $\boldsymbol{H}_t \in \mathcal{R}^{N\times d_h}$ represents the hidden state at time t of the iterative solution, with d_h representing the dimension of the hidden state; $\boldsymbol{x}_t \in \mathcal{R}^{N\times 1}$ represents the observation values of all monitoring stations at time t; $\boldsymbol{Z}_{t-1}$, $\boldsymbol{R}_{t-1}$, $\ddot{\boldsymbol{H}}_{t-1}$, and $\tilde{\boldsymbol{H}}_{t-1}$ represent temporary variables during the iteration process; σ represents the sigmoid activation function; tanh represents the hyperbolic tangent activation function; $\odot$ represents the hadamard product; $[\cdot||\cdot]$ represents the concatenate function; $g\left(\boldsymbol{H},\tau,\{i\}_{i=1}^{N}\right)$ represents the spatio-temporal derivative network; Similar to NODE, we use a numerical ODE solver to implement the solving process of ordinary differential equations (Chen *et al.* 2018); $\Phi_{\boldsymbol{W}_{ic}}*$ represents the convolution operation for the residual connection, and $\boldsymbol{W}_{ic} \in \mathcal{R}^{d_h\times k\times 1\times 1}$ represents the convolution kernel of the convolutional network; and $\boldsymbol{W}_z \in \mathcal{R}^{N\times(d_h+1)}$, $\boldsymbol{W}_r \in \mathcal{R}^{N\times(d_h+1)}$, and $\boldsymbol{W}_h \in \mathcal{R}^{N\times(d_h+1)}$ denotes the weight of the hidden state in the iterative solution process.

4.1.3. Forward propagation of the STA

After obtaining the hidden states $\boldsymbol{\mathcal{H}}$ through the ST-ODE module, we aim to capture the long-term temporal dependencies in $\boldsymbol{\mathcal{H}}$ to improve the prediction capability of the model. To ensure the interpretability of the model, we propose a spatio-temporal attention module to fuse k hidden states for capturing the long-term temporal dependence in data because existing studies have shown that the attention mechanism is an interpretable data structure (Samek *et al.* 2021). Specifically, the spatio-temporal attention module consists of multiple spatial attention blocks and multiple temporal attention blocks. The spatial attention block fuses k hidden states in the spatial dimension, while the temporal attention block is used to fuse k hidden states in the temporal dimension.

In contrast to the calculation method of the traditional attention mechanism, we adopt a more convenient calculation method, which is inspired by Guo *et al.* (2019). Figure 6 illustrate the fusion process of the spatio-temporal attention module, which consists of a single-spatial attention block and a single-temporal attention block. In the spatio-temporal attention module, we also use a residual connection to improve the optimization efficiency of the model. The computation process of Figure 6 is described in Equations (7)–(9).

$$\begin{cases} \hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}} = \Phi_{\boldsymbol{W}_o^{\mathcal{T}}}*\tilde{\boldsymbol{\mathcal{H}}}^{\mathcal{T}} \\ \tilde{\boldsymbol{\mathcal{H}}}^{\mathcal{T}} = \Phi_{\boldsymbol{W}_{rc}^{\mathcal{T}}}*\boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{H}}\boldsymbol{A}^{\mathcal{T}} \\ \boldsymbol{A}_{ij}^{\mathcal{T}} = \dfrac{\exp(\tilde{\boldsymbol{A}}_{ij}^{\mathcal{T}})}{\sum_{j=1}^{n}\exp(\tilde{\boldsymbol{A}}_{ij}^{\mathcal{T}})} \\ \tilde{\boldsymbol{A}}^{\mathcal{T}} = \left(\boldsymbol{\mathcal{H}}^{T}\boldsymbol{W}_{Q}^{\mathcal{T}}\right)\boldsymbol{W}_{K}^{\mathcal{T}}\left(\boldsymbol{W}_{V}^{\mathcal{T}}\boldsymbol{\mathcal{H}}\right) \end{cases} \tag{7}$$

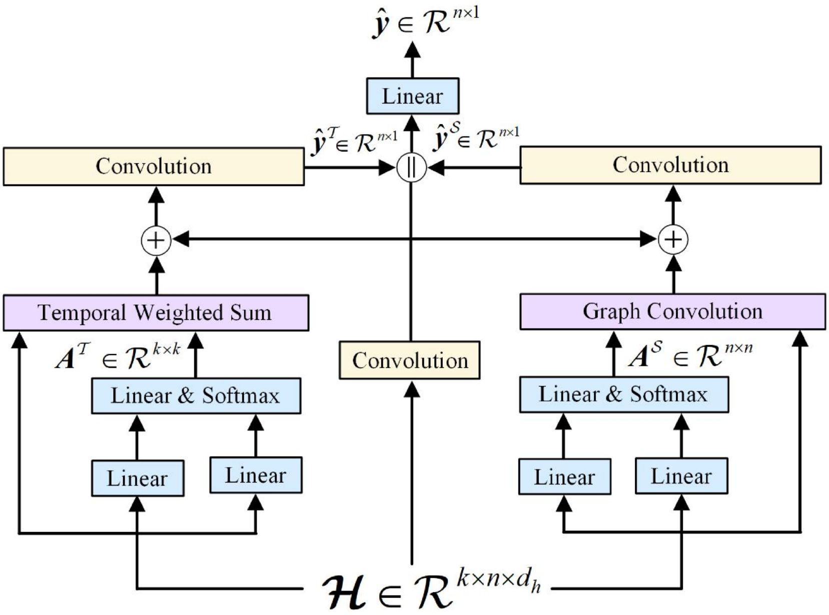
$\hat{y} \in \mathcal{R}^{n\times1}$
Linear
$\hat{y}^{\mathcal{T}} \in \mathcal{R}^{n\times1}$
$\hat{y}^{\mathcal{S}} \in \mathcal{R}^{n\times1}$
Convolution
Convolution
Temporal Weighted Sum
Graph Convolution
$A^{\mathcal{T}} \in \mathcal{R}^{k\times k}$
Convolution
$A^{\mathcal{S}} \in \mathcal{R}^{n\times n}$
Linear & Softmax
Linear & Softmax
Linear
Linear
Linear
Linear
$\mathcal{H} \in \mathcal{R}^{k\times n\times d_h}$

Figure 6. Forward propagation of the STA module: the right side represents the fusion of *k* hidden states from the spatial dimension, and the left side represents the fusion of *k* hidden states from the temporal dimension.

$$\begin{cases} \hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} = \Phi_{\boldsymbol{W}_o^{\mathcal{S}}} * \tilde{\mathcal{H}}^{\mathcal{S}} \\ \tilde{\mathcal{H}}^{\mathcal{S}} = \Phi_{\boldsymbol{W}_{rc}^{\mathcal{S}}} * \mathcal{H} + \left(\mathcal{H}^{T}\boldsymbol{A}^{\mathcal{S}}\right)^{T} \\ \boldsymbol{A}_{ij}^{\mathcal{S}} = \dfrac{\exp(\tilde{\boldsymbol{A}}_{ij}^{\mathcal{S}})}{\sum_{j=1}^{n}\exp(\tilde{\boldsymbol{A}}_{ij}^{\mathcal{S}})} \\ \tilde{\boldsymbol{A}}^{\mathcal{S}} = \left(\mathcal{H}\boldsymbol{W}_Q^{\mathcal{S}}\right)\boldsymbol{W}_K^{\mathcal{S}}\left(\boldsymbol{W}_V^{\mathcal{S}}\mathcal{H}\right)^{T} \end{cases} \tag{8}$$

$$\hat{\boldsymbol{x}}_{t+1} = \left[\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} \middle\| \hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}}\right]\boldsymbol{W}_o \tag{9}$$

where $\mathcal{H} \in \mathcal{R}^{N\times d_h\times k}$ represents *k* hidden states obtained through Section 4.1.2; $\mathcal{H}^{T} \in \mathcal{R}^{k\times d_h\times N}$ represents the transposition of $\mathcal{H}$; $\hat{\boldsymbol{x}}_{t+1} \in \mathcal{R}^{N\times1}$ represents the predicted results at the $(t+1)$*th* time window; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} \in \mathcal{R}^{N\times1}$ represents the fusion values of the spatial dimension at the $(t+1)$*th* time window; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}} \in \mathcal{R}^{N\times1}$ represents the fusion values of the temporal dimension at the $(t+1)$*th* time window; $\boldsymbol{A}^{\mathcal{S}} \in \mathcal{R}^{N\times N}$ represents the spatial attention matrix, which represents the correlation relationships between nodes; $\mathcal{H}^{T}\boldsymbol{A}^{\mathcal{S}} \in \mathcal{R}^{k\times d_h\times N}$ represents a weighted sum of spatial dimensions, which is equivalent to the graph convolution operation of the graph attention. Since we establish a fully connected graph structure, the mathematical expression of the weighted sum in spatial dimensions is the same as that of the graph attention mechanism (Veličković *et al.* 2018); $\boldsymbol{A}^{\mathcal{T}} \in \mathcal{R}^{k\times k}$ represents the temporal attention matrix; $\mathcal{H}\boldsymbol{A}^{\mathcal{T}} \in \mathcal{R}^{N\times d_h\times k}$ represents the weighted sum of temporal dimensions; $\boldsymbol{W}_Q^{\mathcal{T}} \in \mathcal{R}^{N}$, $\boldsymbol{W}_K^{\mathcal{T}} \in \mathcal{R}^{d_n\times N}$, $\boldsymbol{W}_V^{\mathcal{T}} \in \mathcal{R}^{d_n}$, $\boldsymbol{W}_Q^{\mathcal{S}} \in \mathcal{R}^{k}$, $\boldsymbol{W}_K^{\mathcal{S}} \in \mathcal{R}^{d_n\times k}$, $\boldsymbol{W}_V^{\mathcal{S}} \in \mathcal{R}^{d_n}$, and $\boldsymbol{W}_o \in \mathcal{R}^{2\times1}$ represent learnable parameters in the fully connection layer; $\Phi_{\boldsymbol{W}_o^{\mathcal{T}}}*$, $\Phi_{\boldsymbol{W}_{rc}^{\mathcal{T}}}*$, $\Phi_{\boldsymbol{W}_o^{\mathcal{S}}}*$, and $\Phi_{\boldsymbol{W}_{rc}^{\mathcal{S}}}*$

represent the convolution operation, where $\Phi_{\boldsymbol{W}_o^{\mathcal{T}}}*$ and $\Phi_{\boldsymbol{W}_o^{\mathcal{S}}}*$ are used for dimensional alignment, and $\Phi_{\boldsymbol{W}_{rc}^{\mathcal{T}}}$ and $\Phi_{\boldsymbol{W}_{rc}^{\mathcal{S}}}*$ are used for residual connection; $\boldsymbol{W}_o^{\mathcal{T}} \in \mathcal{R}^{N\times N\times d_h\times k}$, $\boldsymbol{W}_{rc}^{\mathcal{T}} \in \mathcal{R}^{N\times N\times 1\times 1}$, $\boldsymbol{W}_o^{\mathcal{S}} \in \mathcal{R}^{N\times N\times d_h\times k}$, and $\boldsymbol{W}_{rc}^{\mathcal{S}} \in \mathcal{R}^{N\times N\times 1\times 1}$ represent the learnable parameters in the convolution layer; $\tilde{\boldsymbol{\mathcal{H}}}^{\mathcal{S}} \in \mathcal{R}^{N\times d_h\times k}$, $\tilde{\boldsymbol{\mathcal{H}}}^{\mathcal{T}} \in \mathcal{R}^{N\times d_h\times k}$, $\tilde{\boldsymbol{A}}^{\mathcal{T}} \in \mathcal{R}^{N\times N}$, and $\tilde{\boldsymbol{A}}^{\mathcal{S}} \in \mathcal{R}^{N\times N}$ represent temporary variables; and exp stands for the exponential function.

Compared to iterative models such as RNN and GRU, which depend solely on the nearest hidden state, the proposed STA-ODE model can explicitly establish the correlation between the prediction value and k hidden states in the time dimension through Equation (7). This allows the proposed STA-ODE model to effectively capture long-term time dependencies, differing from iterative models such as RNN and GRU. Similarly, the proposed STA-ODE model can explicitly establish the correlation between the target node and all nodes in the spatial dimension through Equation (8). Since the proposed STA-ODE model establishes a correlation between the target node and all nodes (even if the nodes are far away from the target node), it can capture long-range spatial dependencies.

4.2. Optimization of the STA-ODE

During the forward propagation, the STA-ODE model predicts future spatio-temporal data $\hat{\boldsymbol{x}}_{t+1}$ based on historical k spatio-temporal data $\boldsymbol{X}_{t-k+1}^{t} = \{\boldsymbol{x}_{t-k+1}, \boldsymbol{x}_{t-k+2}, \ldots\ldots, \boldsymbol{x}_t\}$. Theoretically, the final prediction model can be trained by minimizing the square loss between the ground truth $\boldsymbol{x}_{t+1}$ and the prediction value $\hat{\boldsymbol{x}}_{t+1}$. However, merely optimizing the square loss between $\hat{\boldsymbol{x}}_{t+1}$ and $\boldsymbol{x}_{t+1}$ ignores the alignment of the fusion results in the temporal and spatial dimensions. When the fusion results of a single dimension (ie either $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}}$ or $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}}$) significantly deviate from the ground truth, we may observe inferior prediction performance due to the accumulation of errors. In practical scenarios, both the fusion results $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}}$ of the spatial dimension and the fusion results $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}}$ of the temporal dimension describe the inherent characteristics of spatio-temporal data. Therefore the fusion result of each dimension should be as close to the ground truth as possible. Therefore, inspired by multi-view learning (Cheng *et al.* 2019, Wang *et al.* 2022b), we integrate the alignment of the fusion results into the optimization process of the STA-ODE, and the corresponding loss function is shown in Equation (10).

$$\mathcal{L}(\boldsymbol{W}) = \min_{\boldsymbol{W}} \left(||\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}_{t+1}||_2^2 + \alpha||\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}} - \boldsymbol{x}_{t+1}||_2^2 + \beta||\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} - \boldsymbol{x}_{t+1}||_2^2 \right) \tag{10}$$

where $\boldsymbol{x}_{t+1} \in \mathcal{R}^{N\times 1}$ represents the ground truth of the $(t+1)th$ time window; $\hat{\boldsymbol{x}}_{t+1} \in \mathcal{R}^{N\times 1}$ represents the prediction values of the $(t+1)th$ time window; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{T}} \in \mathcal{R}^{N\times 1}$ represents the fusion values of the temporal dimension for the $(t+1)th$ time window; $\hat{\boldsymbol{x}}_{t+1}^{\mathcal{S}} \in \mathcal{R}^{N\times 1}$ represents the fusion values of the spatial dimension for the $(t+1)th$ time window; $||\cdot||_2^2$ represents a function that solves the 2-norm of a vector; $\boldsymbol{W}$ indicates the learnable parameters in the STA-ODE model; α and β represent the regularization terms that penalize the deviation between the fusion results and the ground truth. In generally, when α is greater (less) than β, it indicates that the impact of temporal correlation on the prediction results is greater (less) than that of spatial correlation. When α is equal to β, it indicates that the impact of temporal correlation on the prediction results is equal to the impact of spatial correlation.

5. Experimental results and discussions

5.1. Data preparation

5.1.1. Data sources

Three spatio-temporal datasets were used to evaluate the performance of the STA-ODE model, namely, traffic volume data, PM2.5 monitoring data, and temperature monitoring data. Table 1 shows the statistical characteristics of the three spatio-temporal datasets.

The traffic volume dataset comes from 67 monitoring cameras in Wuhan, China (Wang *et al*. 2023). Figure 7(a) shows the spatial distribution of monitoring cameras. The time span of the traffic volume dataset is from March 01, 2021, to March 28, 2021, and the time window size is 5 minutes. Each traffic volume data record contains the unique identification of the monitoring camera, the coordinates of the monitoring camera, the monitoring time window, and the traffic volume within the time window.

The PM2.5 monitoring dataset comes from 36 air quality monitoring stations in Beijing, China (Zheng *et al*. 2015). Figure 7(b) shows the spatial distribution of the air quality monitoring stations. The time span of the PM2.5 monitoring dataset is from May 1, 2014, to August 31, 2014, and the time window size is 60 min. Each PM2.5 data record contains the unique identification of the monitoring station, the coordinates of the monitoring station, the monitoring time window, and the PM2.5 air content within the time window.

Table 1. Description of the datasets.

Dataset	Traffic volume	PM2.5	Temperature
Location	Wuhan	Beijing	Wuhan and its surrounding areas
Size of time window	5 min	60 min	60 min
Number of spatial objects	67	36	64
Number of temporal objects	8064	2952	2208
Time span	2021/3/1–2021/3/28	2014/5/1–2014/8/31	2018/6/1–2018/8/31

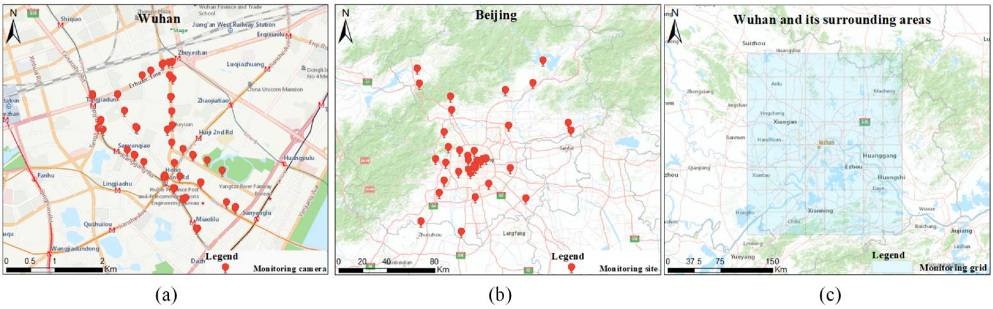

Figure 7. Study area: (a) monitoring cameras in the traffic flow dataset, (b) monitoring sites in the PM2.5 dataset, and (c) experimental grids in the temperature dataset.

The temperature monitoring dataset comes from the Copernicus climate database (Hersbach et al. 2018), and records the air temperature at 2 meters above the surface of inland waters (the time window size is 60 min). As shown in Figure 7(c), we selected 64 0.25°×0.25° grids in Wuhan and its surrounding areas for the experiment. Each temperature data record contains the unique identification of the grid, the center point coordinates of the grid, the monitoring time window, and the average temperature in the time window.

5.1.2. Data preprocessing

To support the research of this work, we further pre-processed three spatio-temporal datasets, and the pre-processing process is described as follows:

1. Due to the limitations of collection technologies and privacy issues, the collected spatio-temporal data may be naturally missing, which may affect the prediction performance of the subsequent model. Therefore, we used the BTTF model to impute the naturally missing values.
2. The processed datasets (ie spatio-temporal states in Definition 2) were manually divided into training and test samples. According to the 20–80 criterion, the training samples account for 80%, and the test samples account for 20%.

5.2. Evaluation metrics

In spatio-temporal prediction, a key problem is how to evaluate the prediction performance of the model. In this study, mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) are used as quantitative indicators to verify the prediction accuracy of the proposed model. The calculation methods of MAE, RMSE, and MAPE are shown in Equations (11)–(13), respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_{t+1}^{i} - \hat{x}_{t+1}^{i}\right| \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{t+1}^{i} - \hat{x}_{t+1}^{i}\right)^{2}} \tag{12}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{x_{t+1}^{i} - \hat{x}_{t+1}^{i}}{x_{t+1}^{i}}\right| \tag{13}$$

where x_{t+1}^{i} represents the ground truth of node v_i in the $t+1$ time window; $\hat{x}_{t+1}^{i}$ represents the predicted value of node v_i in the $t+1$ time window; N represents the total number of nodes in the study area.

5.3. Experimental settings

In this subsection, we describe the experimental environment (hardware and software environment) and the hyper-parameter setting information.

In this study, the spatio-temporal data was processed on a PC (CPU: Intel(R) Xeon(R) E-2224G @ 3.50 GHz, memory: 16.0GB). We built our model based on PyTorch

Table 2. Parameter setting (in Traffic/PM2.5/Temperature) of STA-ODE model.

Hyper-parameter	Range	Optimal value
Time dependent step (k)	[1,2,3, … ,10]	10/6/4
Hidden state dimension (d_h)	[64,128,192,256]	128/192/128
TA block number (n_b^{ta})	[1,2,3,4]	3/2/1
SA block number (n_b^{sa})	[1,2,3,4]	3/2/1
Regularization coefficient (α)	[1,2,3,4,5,6]	5/4/3
Regularization coefficient (β)	[1,2,3,4,5,6]	3/2/3

and Python3.7 on a Graphics Processing Unit (GPU) platform with 24GB of GPU memory.

The hyper-parameters of the STA-ODE model mainly include the time-dependent step k, the hidden state dimension d_h, the TA block number n_b^{ta}, SA block number n_b^{sa}, the regularization coefficient α, and the regularization coefficient β. Similar to the work of Cheng *et al.* (2020), we used the control variable method to obtain the optimal combination of hyper-parameters. Specifically, the optimal parameter can be obtained by observing the change curve of the quantitative index, ie the optimal parameter is the corresponding value when the error is the smallest. Table 2 shows the search range of the hyper-parameters and the optimal value of the hyper-parameters in the three spatio-temporal datasets. The results indicate that the impact of temporal correlation on the prediction results is greater than impact of spatial correlation in the traffic and PM2.5 datasets (ie $\alpha > \beta$). In the temperature dataset, the impact of temporal correlation on the prediction results is approximately equal to the impact of spatial correlation (ie $\alpha \approx \beta$), which is consistent with common sense that the changes in temperature have strong patterns in both temporal and spatial dimensions.

5.4. Comparison with baselines

Given that knowledge-driven models often have low prediction performance compared to data-driven models, we mainly compared the STA-ODE model with data-driven methods. The baselines used in this study can be roughly divided into two categories. The first category includes the T-GCN (Zhao *et al.* 2020), BiSTGN (Wang *et al.* 2022b), ASTGCN (Guo *et al.* 2019), and DSTAGNN methods (Lan *et al.* 2022), which are regarded as black-box data-driven models. The second category includes the Latent-ODEs (Chen *et al.* 2018), ODE-RNNs (Rubanova *et al.* 2019), and STGODE methods (Fang *et al.* 2021), which are regarded as NODE-based models.

Table 3 compares the STA-ODE model and the baselines in the three datasets (the metrics are the average results of five different initial seeds). The results indicate that the prediction accuracy of the first-category models has considerably improved in recent years. The DSTAGNN model has attained state-of-the-art prediction accuracy in first-category models. Compared to the first-category models, the prediction accuracy of the second-category models has improved slightly in recent years. Among them, the prediction accuracy of the STGODE model is slightly higher than that of the ODE-RNNs model, which in turn is slightly better than that of the Latent-ODEs model. Overall, the prediction accuracy of the second-category models is slightly lower than that of the first-category models. The main reason for the above results is that the NODE-based models enhance the interpretability of the model while sacrificing

Table 3. Comparison results (in MAE/RMSE/MAPE) of prediction performance between STA-ODE and baselines.

Model	Traffic volume	PM2.5	Temperature
T-GCN	5.36/9.06/35.54%	10.89/15.65/31.56%	0.88/1.20/3.16%
BiSTGN	4.42/7.36/26.31%	9.35/14.28/26.78%	0.62/0.99/2.19%
ASTGCN	4.03/6.52/24.87%	7.90/12.07/22.09%	0.58/0.89/2.09%
DSTAGNN	3.97/6.40/22.95%	7.78/12.04/22.27%	0.59/0.84/2.08%
Latent-ODEs	4.17/6.69/25.83%	8.25/13.05/24.19%	0.69/1.04/2.44%
ODE-RNNs	4.16/6.65/25.02%	8.11/12.88/23.01%	0.65/1.01/2.27%
STGODE	4.10/6.87/25.28%	8.08/12.52/22.46%	0.62/0.95/2.19%
STA-ODE	**3.88/6.26/22.87%**	**7.59/11.59/21.18%**	**0.57/0.76/2.06%**

Metrics are the average results of five different initial seeds.

Table 4. The results of statistical significance tests based on average MAE.

	Traffic Volume		PM2.5		Temperature	
Model	*t*-statistic	*p*-value	*t*-statistic	*p*-value	*t*-statistic	*p*-value
T-GCN	−20.63	3.19×10^{-8}	−15.07	3.70×10^{-7}	−11.01	4.11×10^{-6}
BiSTGN	−8.32	3.26×10^{-5}	−11.48	2.98×10^{-6}	−3.39	9.47×10^{-3}
ASTGCN	−4.05	3.63×10^{-3}	−5.99	3.25×10^{-4}	−1.30	**2.27×10^{-1}**
DSTAGNN	−2.20	**5.86×10^{-2}**	−4.73	1.47×10^{-3}	−0.96	**3.61×10^{-1}**
Latent-ODEs	−5.74	4.29×10^{-4}	−9.94	8.83×10^{-6}	−5.40	6.41×10^{-4}
ODE-RNNs	−4.86	1.24×10^{-3}	−6.67	1.56×10^{-4}	−7.95	4.54×10^{-5}
STGODE	−5.33	6.96×10^{-4}	−7.58	6.37×10^{-5}	−4.86	1.25×10^{-3}

Bold indicates that the experimental results did not pass the hypothesis test with a significance level of 5%.

prediction accuracy. Specifically, mining spatial dependencies in spatio-temporal data remains challenging in the above NODE-based models, and the above NODE-based models make it difficult to extract long-term temporal dependencies in data. Compared to baselines, the STA-ODE model mitigates the two shortcomings mentioned above. The STA-ODE model can discover not only spatial dependencies in data, but also long-term temporal dependencies in data, leading to the proposed STA-ODE can obtain better prediction accuracy. There are variations in the prediction accuracy of the STA-ODE model across the three datasets. For example, t the temperature dataset yields higher accuracy compared to the traffic volume dataset and PM2.5 datasets. The main reason is that temperature data move smoothly in space and time, making it easier to predict temperature patterns.

In addition, we conducted a hypothesis test to further the statistical differences in prediction accuracy. In the hypothesis test, we assumed that the evaluation index follows a normal or approximately normal distribution, and then used the *t*-test to analyze whether the means of the two evaluation indicators were equal in the case of unknown variance. Table 4 presents the results of the statistical significance test between the STA-ODE and baselines for MAE. The results show that, in most cases, the STA-ODE yielded significant improvements compared to baselines. Only three groups out of a total of 108 tests) failed the hypothesis test at a 5% significance level. Notably, there was no significant difference in the prediction accuracy between the STA-ODE and DSTAGNN models in the traffic dataset. In the temperature dataset, there is no significant difference in prediction accuracy among the STA-ODE model, the DSTAGNN model, and the ASTGAN model. The above results may be an error caused by the small number of five initial seeds. Furthermore, even though STA-ODE does not have a significant advantage in prediction accuracy, STA-ODE still has certain advantages over DSTAGNN and ASTGAN in terms of model interpretability (please see Section 5.8 for more details).

It is worth mentioning that we only evaluated the prediction accuracy of the STA-ODE model for single-step predictions, it has the capability to make multi-step predictions. The main reason is that iterative models such as STA-ODE can use the output values from single-step predictions as input to generate long-term prediction results. Many studies have indicated that the accuracy of the multi-step prediction depends on the accuracy of single-step prediction (Zhao *et al*. 2020, Wang *et al*. 2023). When the prediction accuracy of single-step prediction is higher, the accuracy of multi-step prediction will also be higher (Chen and Sun 2022, Ji *et al*. 2022). Therefore, Therefore, the proposed STA-ODE model will also have high accuracy in multi-step prediction, as its performance in single-step prediction is superior.

5.5. Qualitative analysis of prediction results

This section utilizes line charts and maps to qualitatively depict the prediction performance of the STA-ODE model. Figure 8 illustrates the discrepancy between the predicted value and ground truth across the temporal dimension. The results indicate that among all three data sets, the residuals between ground truth and predicted values are minor for most times (only occasionally significant, as marked by a blue circle). The main reason why the residuals of the STA-ODE model are large in the blue circle is that the trend of spatio-temporal observations has changed suddenly within a short period. An instance of this can be found in the traffic dataset, where the period with the most significant residual errors is typically during rush hours. Predicting the unexpected increase in traffic volume during rush hours is problematic.

Figure 9 shows the difference between the predicted value and ground truth from the spatial dimension. Similar to the residuals in the temporal dimension, the residuals for most monitoring stations are small, and only a few monitoring stations have significant residuals (marked by a blue circle). Within the traffic dataset, the areas with large residuals are mainly situated on the main road. The model's accuracy suffers due to the main road's highly variable traffic volume. An analysis of the PM2.5 dataset reveals that the southeast of Beijing has the most significant residuals. The main reason is that the southeast area of Beijing is the primary hub of human activities, which

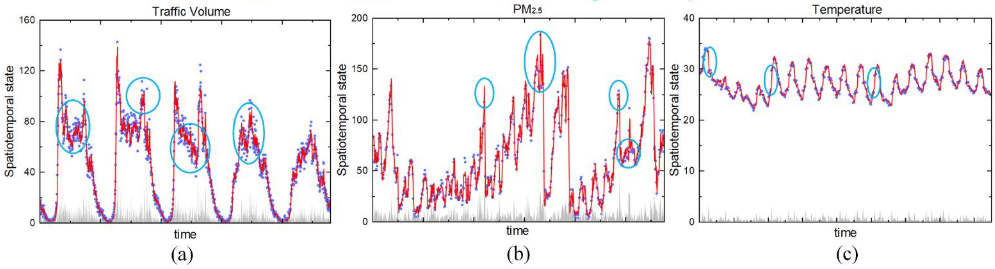

Figure 8. Prediction error of temporal dimension: (a) traffic dataset, (b) pm2.5 dataset, and (c) temperature dataset.

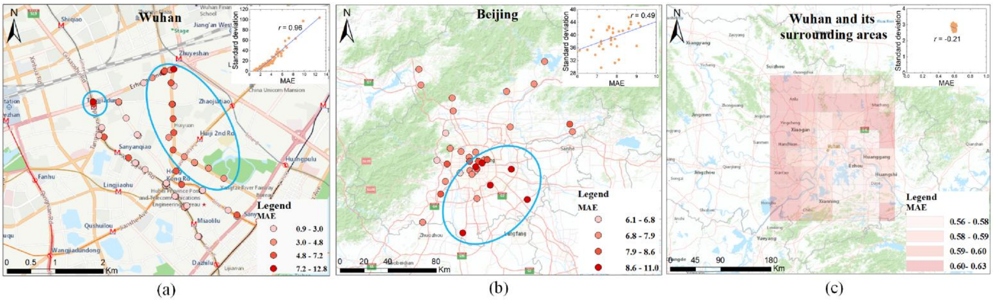

Figure 9. Prediction error of spatial dimension: (a) traffic dataset, (b) pm2.5 dataset, and (c) temperature dataset.

Table 5. Impact of different components on prediction results (in MAE/RMSE/MAPE).

Model	Traffic volume	PM2.5	Temperature
NODE	4.23/7.12/26.27%	8.39/13.07/24.93%	0.73/0.95/2.61%
ST-ODE	4.08/6.73/25.72%	8.02/12.29/23.32%	0.62/0.91/2.20%
ST-ODE/TA	3.94/6.38/23.71%	7.72/11.99/22.47%	0.59/0.86/2.09%
ST-ODE/SA	4.01/6.58/23.83%	7.94/12.39/22.65%	0.61/0.90/2.13%
STA-ODE	**3.88/6.26/22.87%**	**7.59/11.59/21.18%**	**0.57/0.76/2.06%**

leads to significant fluctuation in PM2.5 levels, making accurate predictions difficult. Furthermore, we computed the correlation coefficient between the observed variance and the prediction accuracy. The results indicate a significant positive correlation between the observed variance and the prediction accuracy in both the traffic volume and PM2.5 datasets, demonstrating that considerable fluctuations in observation may lead to low prediction accuracy. Compared to the traffic and PM2.5 datasets, there is no significant correlation between observed variance and prediction accuracy in the temperature dataset. The main reason is that the residual fluctuation of temperature data is relatively small. Moreover, the current results of temperature data do not reflect the relationship between observed variance and predicted values.

Overall, the STA-ODE model demonstrates enhanced prediction accuracy across the temporal and spatial dimensions, effectively capturing trends in spatio-temporal data and proving its good prediction performance.

5.6. Effect of different components on prediction performance

In this subsection, we assess the influence of different components on the prediction results, as displayed in Table 5. NODE represents the classic NODE model, ST-ODE represents the spatio-temporal ordinary differential equation module, ST-ODE/TA represents the STA-ODE method of fusing multiple hidden states with temporal attention only, and ST-ODE/SA represents the STA-ODE method of fusing multiple hidden states with spatial attention only. The results show that the prediction performance of ST-ODE surpasses that of the classic NODE model, thereby indicating that the incorporation of the spatio-temporal derivative network effectively extends the NODE-based time series prediction model to a spatio-temporal prediction model. We also observe that the prediction performance of ST-ODE/TA and ST-ODE/SA is superior to that of

ST-ODE. The results indicate that incorporating temporal and spatial attention can enhance the model's prediction accuracy. The main benefit is that temporal attention can capture long-term temporal dependencies in data, while the spatial attention mechanism can capture long-range spatial dependencies. Furthermore, compared to the prediction results of ST-ODE/TA and ST-ODE/SA, the prediction accuracy of the STA-ODE model has been further improved. The above results indicate that the spatio-temporal attention module can effectively capture spatio-temporal dependencies.

5.7. Effect of loss function on prediction performance

We design a loss function to solve the alignment problem in the fusion results across temporal and spatial dimensions. Therefore, in this subsection, we analyze the impact of the loss function on the prediction performance. Table 6 shows the influence of the loss function on prediction accuracy. Here, STA-ODE-NoAligned refers to the STA-ODE model without alignment in the loss function. The results show that the STA-ODE model has better predictive performance than the STA-ODE-NoAligned model. Especially in the PM2.5 dataset, the prediction accuracy of the STA-ODE model with result alignment has been dramatically improved, proving the effectiveness of loss alignment.

5.8. Analysis of model interpretability

In this subsection, we seek to explain why the STA-ODE model has superior performance. We explain the STA-ODE model from two aspects: one is to delve into the derivative value learned by the spatio-temporal ordinary differential equation module, and the other is to investigate the spatio-temporal relationship learned by the spatio-temporal attention module.

Given that the derivative values of hidden states are challenging to understand, we treat the observed value at each time as solutions to the ordinary differential equation and train a prediction model separately. As shown in Figure 10, we can further infer the trend of spatio-temporal data throughout the day based on the visualized derivative values. When the derivative value exceeds 0, the spatio-temporal data displays an upward trend. When the derivative value falls below 0, the spatio-temporal data exhibits a downward trend. When the derivative value equals 0, the spatio-temporal data remains stable. Taking the traffic dataset as an example, we can explain the traffic changes throughout the day based on the derivative values. Between 0:00 and 7:00, the traffic volume remains relatively stable, as indicated by a derivative value that fluctuates around 0. From 7:00 to 9:00, there is a steady increase in the derivative value of the traffic volume above 0, indicating that the traffic volume continues to grow and reaches its maximum value around 9:00. From 9:00 to 10:00, the derivative value of

Table 6. Impact of loss function on prediction results (in MAE/RMSE/MAPE).

Dataset	STA-ODE-NoAligned	STA-ODE
Traffic Volume	3.94/6.38/23.82%	3.88/6.26/22.87%
PM2.5	7.76/12.063/21.78%	7.59/11.59/21.18%
Temperature	0.58/0.83/2.16%	0.57/0.76/2.06%

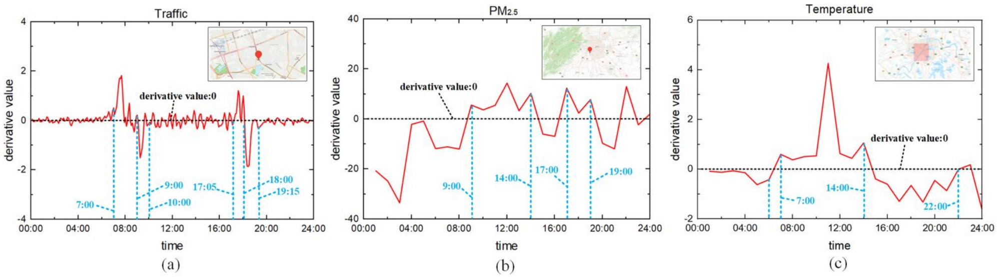

Figure 10. Derivative value learned in ST-ODE: (a) traffic dataset, (b) pm2.5 dataset, and (c) temperature dataset.

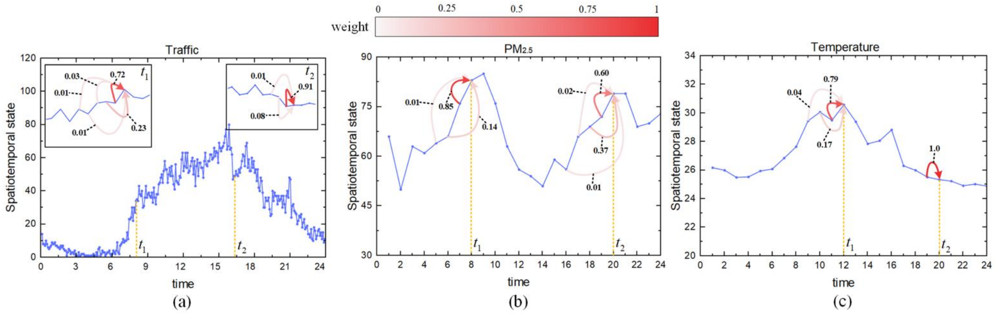

Figure 11. Weight value learned in temporal attention: (a) traffic dataset, (b) pm2.5 dataset, and (c) temperature dataset.

the traffic volume is less than 0, meaning that the traffic volume rapidly decreases and reaches a balanced stage roughly by 10:00. Similarly, in the PM2.5 dataset, the concentration of PM2.5 increases mainly from 9:00 to 14:00 and from 17:00 to 19:00. In the temperature dataset, the primary period for an increase in temperature is between 7:00 and 14:00. The above results are just in line with our common sense, proving that ST-ODE module has good interpretability in capturing the changing trend of spatio-temporal data. In addition, the ST-ODE module accurately identified the derivative of the observation curve in the spatio-temporal data, further justifying the explicit modeling of spatial information in the spatio-temporal derivative network. In other words, it is further proved that the spatio-temporal derivative network successfully extends the NODE-based time series prediction model into a spatio-temporal prediction model.

In addition to the module of the spatio-temporal ordinary differential equation, the spatio-temporal attention module is also an essential interpretable component. Figure 11 illustrates the temporal dependencies learned in the temporal attention module. The proposed model learns the influence weight of historical observations on future predictions and then predicts the future spatio-temporal data. For example, in the traffic dataset, the predicted value at time t_1 is affected by five historical moments, while the predicted value at the time t_2 is affected by three historical moments. Similarly, in the PM2.5 dataset, the predicted value at time t_1 is affected by three historical moments, whilethe predicted value at the time t_2 is affected by four historical

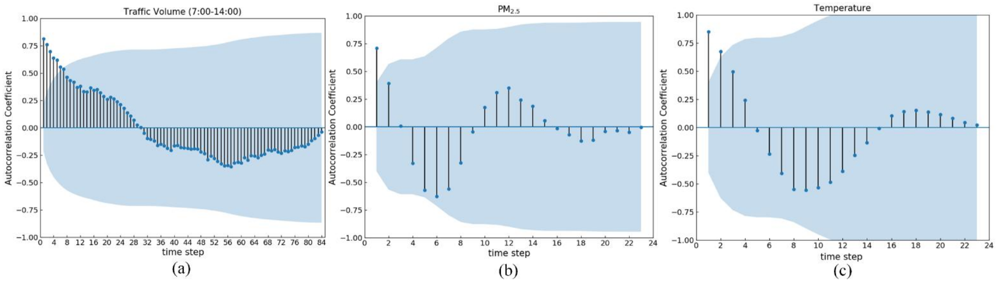

Figure 12. Autocorrelation results in temporal attention: (a) traffic dataset, (b) pm2.5 dataset, and (c) temperature dataset.

moments. In the temperature dataset, the predicted value at the time t_1 is affected by three historical moments, while the predicted value at the time t_2 is affected by one historical moment. The results indicate that the influence weight on the prediction value increases as the historical moment gets closer to the target moment. The visualization results align with our common sense, proving that the spatio-temporal attention module has good interpretability. In addition, the results indicate that the proposed model can capture the weights of five or even more time nodes on the prediction results, proving that the attention module enables the STA-ODE model to capture long-term temporal dependence.

We also conducted a quantitative analysis of the temporal correlation using the autocorrelation function, and the results are shown in Figure 12. Among them, the blue area represents the error range and reveals that the influence of the time step in the area on the prediction results is not significant. The results show that the prediction results are affected by long-term temporal dependence in the traffic dataset. By contrast, the prediction results are affected by short-term temporal dependence in the temperature dataset and the PM2.5 dataset. The above results substantiate why temporal attention can capture longer time dependency in the traffic dataset, demonstrating the rationality of capturing long-term temporal dependence in this study.

6. Conclusions and future work

Interpretable spatio-temporal prediction is a popular research topic in geographic big data mining. However, most existing spatio-temporal prediction models face the challenge of balancing prediction accuracy and interpretability. Therefore, we propose a novel spatio-temporal attentional neural differential equation (STA-ODE) model for interpretable spatio-temporal prediction tasks.

Three real spatio-temporal datasets (traffic volume dataset, PM2.5 monitoring dataset, and temperature monitoring dataset) were used to evaluate the predictive performance of the STA-ODE model. The variable control method was implemented to obtain the optimal parameters for the STA-ODE model. We compared seven existing data-driven baselines, including T-GCN, BiSTGN, ASTGCN, DSTAGNN, Latent-ODEs, ODE-RNNs, and STGODE models. Experimental results showed that STA-ODE outperformed seven existing baselines. The effect of different components and the loss function of STA-ODE on the prediction accuracy was examined, proving that the proposed method is suitable for spatio-temporal prediction. Finally, we visually analyzed the reasons for the superior predictive performance of the STA-ODE model.

The limitations of this study are as follows: (1) The STA-ODE model performs spatio-temporal prediction tasks based on ODE. However, the forward and back propagation of ODE is a time-consuming process, which makes the training of STA-ODE models slow; (2) We only validated the single-step prediction performance of the STA-ODE model, and did not validate the multi-step prediction capability of the STA-ODE model; (3) The STA-ODE model is a general spatio-temporal prediction model, but we use only use three specific spatio-temporal datasets to verify the prediction performance of the proposed model. In response to the above limitations, future work will focus on two aspects. First, we will optimize the iterative process of forward and backward propagation to improve the training efficiency of the STA-ODE model. Second, we will validate the multi-step prediction capability of the STA-ODE model. Finally, multi-source spatio-temporal data will be further collected to evaluate the prediction performance of the STA-ODE model.

Acknowledgements

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project was supported by National Key Research and Development Program of China [Grant No. 2021YFB3900803], National Postdoctoral Innovation Talents Support Program [Grant No. BX20230360], Open funds of the Wuhan University-Huawei Geoinformatics Innovation Laboratory [Grant No. TC20210901025-2023-04], National Natural Science Foundation of China [Grant Nos. 42101423 and 42371470], Special Research Assistant Program of Chinese Academy of Sciences, Innovation Project of LREIS [Grant No. 08R8A092YA].

Notes on contributors

Peixiao Wang is a Postdoctoral Fellow from State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences. He received Ph.D. degree under from State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, and received the M.S. degree from The Academy of Digital China, Fuzhou University. His research topics include spatiotemporal data mining, and spatiotemporal prediction, especially focus on spatiotemporal prediction of transportation systems.

Tong Zhang is a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He received the M.Eng. degree in cartography and geographic information system (GIS) from Wuhan University, Wuhan, China, in 2003, and the Ph.D. degree in geography from San Diego State University, and the

University of California at Santa Barbara in 2007. His research topics include urban computing and machine learning.

Hengcai Zhang is an Associate Professor of State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. He received his Ph.D. degree from Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. He is the member of the Theory and Methodology Committee of the Chinese Association of Geographic Information System, and member of Chinese Branch of ACM SIGSPATIAL. His interests focus on spatial-temporal data mining and 3D-Computing.

Shifen Cheng is an Associate Professor of State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. He received his Ph.D. degree from Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests include spatiotemporal data mining, urban computing and intelligent transportation.

Wangshu Wang is a postdoctoral fellow at the Research Unit Cartography at the Vienna University of Technology. She received her Ph.D. degree from the Vienna University of Technology in 2023. Her research focuses on spatiotemporal data mining and indoor pedestrian navigation.

ORCID

Peixiao Wang http://orcid.org/0000-0002-1209-6340
Tong Zhang http://orcid.org/0000-0002-0683-4669
Hengcai Zhang http://orcid.org/0000-0002-5004-9609
Shifen Cheng http://orcid.org/0000-0002-9553-8318
Wangshu Wang http://orcid.org/0000-0003-2307-155X

Data and codes availability statement

The data and codes that support the findings of this study are available in 'figshare.com' with the identifier https://doi.org/10.6084/m9.figshare.22678153.

References

Aryaputera, A.W., *et al.*, 2015. Very short-term irradiance forecasting at unobserved locations using Spatio-temporal kriging. *Solar Energy*, 122, 1266–1278.

Bartier, P.M., and Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences*, 22 (7), 795–799.(96)00021-0

Chen, R.T.Q., *et al.*, 2018. Neural ordinary differential equations. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6572–6583.

Chen, X., and Sun, L., 2022. Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9), 4659–4673.

Cheng, S., Lu, F., and Peng, P., 2021. Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns. *IEEE Transactions on Intelligent Transportation Systems*, 22 (10), 6365–6383.

Cheng, S., *et al.*, 2018. Short-term traffic forecasting: an adaptive ST-KNN model that considers spatial heterogeneity. *Computers, Environment and Urban Systems*, 71, 186–198.

Cheng, S., *et al.*, 2019. Multi-task and multi-view learning based on particle swarm optimization for short-term traffic forecasting. *Knowledge-Based Systems*, 180, 116–132.

Cheng, S., Peng, P., and Lu, F., 2020. A lightweight ensemble spatiotemporal interpolation model for geospatial data. *International Journal of Geographical Information Science*, 34 (9), 1849–1872.

Ermagun, A., and Levinson, D., 2018. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38 (6), 786–814.

Fang, Z., *et al.*, 2021. Spatial-temporal graph ODE networks for traffic flow forecasting. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 364–373.

Guo, S., *et al.*, 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01), 922–929.

Hersbach, H., *et al.*, 2018. *ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.adbb2d47 [Accessed 15 June 2022].

Huang, W., *et al.*, 2021. An overview of air quality analysis by big data techniques: monitoring, forecasting, and traceability. *Information Fusion*, 75, 28–40.

Janowicz, K., *et al.*, 2020. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34 (4), 625–636.

Ji, J., *et al.*, 2022. STDEN: towards physics-guided neural networks for traffic flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (4), 4048–4056.

Kang, Y., *et al.*, 2022. STICC: a multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical Information Science*, 36 (8), 1518–1549.

Lan, S., *et al.*, 2022. DSTAGNN: dynamic spatial-temporal aware graph neural network for traffic flow forecasting. *Proceedings of the 39th International Conference on Machine Learning*, 11906–11917. https://proceedings.mlr.press/v162/lan22a.html

Lechner, M., and Hasani, R., 2020. Learning long-term dependencies in irregularly-sampled time series (arXiv:2006.04418). *arXiv*,

Li, L., *et al.*, 2014. Fast inverse distance weighting-based spatiotemporal interpolation: A web-based application of interpolating daily fine particulate matter PM2:5 in the contiguous U.S. using parallel programming and k-d tree. *International Journal of Environmental Research and Public Health*, 11 (9), 9101–9141.

Li, M., *et al.*, 2021. Prediction of human activity intensity using the interactions in physical and social spaces through graph convolutional networks. *International Journal of Geographical Information Science*, 35 (12), 2489–2516.

Li, M., *et al.*, 2020. Predicting future locations of moving objects with deep fuzzy-LSTM networks. *Transportmetrica A: Transport Science*, 16 (1), 119–136.

Liu, Y., *et al.*, 2016. Urban water quality prediction based on multi-task multi-view learning. https://www.microsoft.com/en-us/research/publication/urban-water-quality-prediction-based-multi-task-multi-view-learning-2/

Duan, P., *et al.*, 2016. STARIMA-based traffic prediction with time-varying lags. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 1610–1615.

Pesquer, L., Cortés, A., and Pons, X., 2011. Parallel ordinary kriging interpolation incorporating automatic variogram fitting. *Computers & Geosciences*, 37 (4), 464–473.

Rubanova, Y., Chen, R.T.Q., and Duvenaud, D., 2019. Latent ODEs for irregularly-sampled time series. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 5320–5330.

Sagi, O., and Rokach, L., 2020. Explainable decision forest: transforming a decision forest into an interpretable tree. *Information Fusion*, 61, 124–138.

Samek, W., *et al.*, 2021. Explaining deep neural networks and beyond: a review of methods and applications. *Proceedings of the IEEE*, 109 (3), 247–278.

Veličković, P., *et al.*, 2018. Graph attention networks (arXiv:1710.10903). *arXiv*, Wang, P., Zhang, T., and Hu, T., 2022. Traffic condition estimation and data quality assessment for signalized road networks using massive vehicle trajectories. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-022-03892-z

Wang, P., *et al.*, 2022a. A hybrid data-driven framework for spatiotemporal traffic flow data imputation. *IEEE Internet of Things Journal*, 9 (17), 16343–16352.

Wang, P., *et al.*, 2022b. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. *International Journal of Geographical Information Science*, 36 (6), 1231–1257.

Wang, P., *et al.*, 2023. Urban traffic flow prediction: a dynamic temporal graph network considering missing values. *International Journal of Geographical Information Science*, 37 (4), 885–912.

Xie, P., *et al.*, 2020. Urban flow prediction from spatiotemporal data using machine learning: a survey. *Information Fusion*, 59, 1–12.

Xu, L., *et al.*, 2021. Spatiotemporal forecasting in earth system science: methods, uncertainties, predictability and future directions. *Earth-Science Reviews*, 222, 103828.

Yozgatligil, C., *et al.*, 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology*, 112 (1-2), 143–167.

Yu, B., *et al.*, 2016. K-nearest neighbor model for multiple-time-step prediction of short-term traffic condition. *Journal of Transportation Engineering*, 142 (6), 04016018.

Yu, B., Yin, H., and Zhu, Z., 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3634–3640.

Yu, H.-F., Rao, N., and Dhillon, I.S., 2016. Temporal regularized matrix factorization for high-dimensional time series prediction. *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 15.

Zhang, J., Zheng, Y., and Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31 (1), 10735.

Zhang, J., *et al.*, 2020. Flow prediction in spatio-temporal networks based on multitask deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 32 (3), 468–478.

Zhang, K., *et al.*, 2021. Graph attention temporal convolutional network for traffic speed forecasting on road networks. *Transportmetrica B Transport Dynamics*, 9 (1), 153–171.

Zhao, L., *et al.*, 2020. T-GCN: a temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21 (9), 3848–3858.

Zheng, G., *et al.*, 2023. Hybrid deep learning models for traffic prediction in large-scale road networks. *Information Fusion*, 92, 93–114.

Zheng, Y., *et al.*, 2015. Forecasting fine-grained air quality based on big data. *Proceedings of the 21th SIGKDD Conference on Knowledge Discovery and Data Mining*. https://www.microsoft.com/en-us/research/publication/forecasting-fine-grained-air-quality-based-on-big-data/

Zhou, F., *et al.*, 2021. Urban flow prediction with spatial–temporal neural ODEs. *Transportation Research Part C*, 124, 102912.

Zhu, D., *et al.*, 2020a. Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, 34 (4), 735–758.

Zhu, D., *et al.*, 2020b. Understanding place characteristics in geographic contexts through graph convolutional neural networks. *Annals of the American Association of Geographers*, 110 (2), 408–420.